Raymond Lih
304937942

CSM146 - Problem Set 3

1. a) By Mercer's Thm, $k(\cdot,\cdot)$ is a kernel function iff for all n, the matrix

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_n) & \cdots & k(x_n, x_n) \end{bmatrix}$$ is positive semidefinite

- A matrix, A, is positive semi-definite iff

$$z^T A z = \sum_{ij} A_{ij} z_i z_j \geq 0 \quad \text{for all } z \in \mathbb{R}^D$$

- so $k(\cdot,\cdot)$ is a kernel function if

$$z^T K z = \sum_{ij} K_{ij} z_i z_j \geq 0 \quad \text{for all } z \in \mathbb{R}^D$$

- we know $k(x_i, x_j) = k(x_j, x_i)$ because the intersection of 2 sets is the same
- thus, K is a symmetric matrix
- thus, K can be decomposed into the product of 2 matrices: $K = A^T A$
- we have

$$z^T K z = z^T A^T A z = (Az)^T (Az)$$

$$(Az)^T (Az) = \sum_{ij} K_{ij} z_i z_j \geq 0$$

any matrix multiplied with its own transpose is positive, because it is basically squaring all its values

- since $A^T A$ positive semidefinite, K is positive semi definite
- each entry of K is $k(\cdot,\cdot)$
- therefore, by Mercer's theorem, $k(\cdot,\cdot)$ is a kernel function □

b) Prove that

$$\left(1+\left(\frac{x}{\|x\|}\right)\left(\frac{z}{\|z\|}\right)\right)^3$$

is a kernel function.

- we know $k(x, z) = x \cdot z$ is a kernel function
- let $k_1(x, z) = \left(\frac{x}{\|x\|}\right)\left(\frac{z}{\|z\|}\right)$

$$= \frac{1}{\|x\|} \cdot x \cdot z \cdot \frac{1}{\|z\|}$$

$$= f(x) \cdot x \cdot z \cdot f(z)$$

where $f(x) = \frac{1}{\|x\|} = \frac{1}{\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}}$

$$k_1(x, z) = f(x) \cdot x \cdot z \cdot f(z)$$

$$= f(x) \cdot k(x, z) \cdot f(z)$$

- thus $k_1(x, z)$ is also a kernel function by the scaling rule
- we have

$$(1 + k_1(x, z))^3$$

- let $k_2(x, z) = 1$
- then $\underline{K_2} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$

(capital $\underline{K}$)

- $\underline{K_2}$ is a symmetric matrix, so it can be decomposed into $K_2 = A^T A$
- then K_2, is positive semidefinite because

$$z^T K_2 z$$

$$= z^T A^T A z$$

$$= (Az)^T (Az)$$

$$= \sum_{ij} A_{ij} z_i z_j \geq 0$$

- thus $k_2(x, z) = 1$ is also a kernel function
- we have
$$(k_2(x, z) + k_1(x, z))^3$$
  - let $k_3(x, z) = k_2(x, z) + k_1(x, z)$
    - k_3 is also a kernel function by the sum rule
- we have
$$(k_3(x, z))^3$$
  - let $k_4(x, z) = (k_3(x, z))^3$
$$= k_3(x, z) \cdot k_3(x, z) \cdot k_3(x, z)$$
  - k_4 is also a kernel function by the product rule
  - $k_4 = (k_3(x, z))^3$
$$= (k_2(x, z) + k_1(x, z))^3$$
$$= \left(1 + \left(\frac{x}{\|x\|}\right) \cdot \left(\frac{z}{\|z\|}\right)\right)^3$$
  - therefore $\left(1 + \left(\frac{x}{\|x\|}\right) \cdot \left(\frac{z}{\|z\|}\right)\right)^3$ is a kernel function

$\square$

c) $K(x, z) = (1 + xz)^3 = \phi(x)^T \phi(z)$

$$= (1 + x_1 z_1 + x_2 z_2)^3$$
$$= (1 + x_1 z_1 + x_2 z_2)(1 + x_1 z_1 + x_2 z_2)^2$$
$$= (1 + x_1 z_1 + x_2 z_2 + x_1 z_1 + x_1^2 z_1^2 + x_1 x_2 z_1 z_2 + x_2 z_2 + x_1 x_2 z_1 z_2 + x_2^2 z_2^2)$$
$$= (1 + 2x_1 z_1 + 2x_2 z_2 + 2x_1 x_2 z_1 z_2 + x_1^2 z_1^2 + x_2^2 z_2^2)(1 + x_1 z_1 + x_2 z_2)$$
$$= 1 + 2x_1 z_1 + 2x_2 z_2 + 2x_1 x_2 z_1 z_2 + x_1^2 z_1^2 + x_2^2 z_2^2$$
$$+ x_1 z_1 + 2x_1^2 z_1^2 + 2x_1 x_2 z_1 z_2 + 2x_1^2 x_2 z_1^2 z_2$$
$$+ x_1^3 z_1^3 + \boxed{x_1 x_2^2 z_1 z_2^2} + x_2 z_2 + 2x_1 x_2 z_1 z_2$$
$$+ 2x_2^2 z_2^2 + \boxed{2x_1 x_2^2 z_1 z_2^2} + x_1^2 x_2 z_1^2 z_2 + x_2^3 z_2^3)$$

$$= 1 + 3x_1 z_1 + 3x_2 z_2 + 3x_1^2 z_1^2 + 3x_2^2 z_2^2 + 6x_1 x_2 z_1 z_2 + 3x_1^2 x_2 z_1^2 z_2 + 3x_1 x_2^2 z_1 z_2^2 + x_1^3 z_1^3 + x_2^3 z_2^3$$

$$\Phi(x) = \begin{bmatrix} 1 \\ \sqrt{3}\, x_1 \\ \sqrt{3}\, x_2 \\ \sqrt{3}\, x_1^2 \\ \sqrt{3}\, x_2^2 \\ \sqrt{6}\, x_1 x_2 \\ \sqrt{3}\, x_1^2 x_2 \\ \sqrt{3}\, x_1 x_2^2 \\ x_1^3 \\ x_2^3 \end{bmatrix} \qquad \Phi_\beta(x) = \begin{bmatrix} 1 \\ \sqrt{3\beta}\, x_1 \\ \sqrt{3\beta}\, x_2 \\ \sqrt{3\beta}\, x_1^2 \\ \sqrt{3\beta}\, x_2^2 \\ \sqrt{6\beta}\, x_1 x_2 \\ \sqrt{3\beta}\, x_1^2 x_2 \\ \sqrt{3\beta}\, x_1 x_2^2 \\ \sqrt{\beta^3}\, x_1^3 \\ \sqrt{\beta^3}\, x_2^3 \end{bmatrix}$$

(.)

The role of the parameter β is to scale the transformation. We can see that most entries to the transformed feature vector are scaled by a factor of $\sqrt{\beta}$, while two of the entries are scaled by a factor of $\sqrt{\beta^3}$.

(1

optimization $(f(\theta))$ constraint $(g(\theta))$

2. $\frac{1}{2}\|\theta\|^2$, $y_n \theta^T x_n \geq 1$, $n = 1 \dots N$

a) $\vec{x}_n = \begin{bmatrix} a \\ e \end{bmatrix}$, $y_n = 1$, $c = 1$

- Lagrangian

$$\mathcal{L}(x, y \dots \lambda) = f(x, y \dots) - \lambda(g(x, y \dots) - c)$$
$$\mathcal{L}(\theta, \lambda) = f(\theta) - \lambda(g(\theta) - C)$$
$$= -\frac{1}{2}\|\theta\|^2 - \lambda(y_n \theta^T x_n - c)$$
$$= \frac{1}{2}\|\theta\|^2 - \lambda(-\theta^T x_n - 1)$$
$$= \frac{1}{2}\|\theta\|^2 + \lambda\theta^T x_n + \lambda$$

$$\nabla \mathcal{L}(\theta, \lambda) = \vec{0} = \begin{bmatrix} \frac{\partial}{\partial\theta_1}\mathcal{L} \\ \frac{\partial}{\partial\theta_2}\mathcal{L} \\ \frac{\partial}{\partial\lambda}\mathcal{L} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathcal{L}(\theta, \lambda) = \frac{1}{2}(\theta_1^2 + \theta_2^2) + \lambda(\theta_1 x_1 + \theta_2 x_2) + \lambda$$

$$\frac{\partial}{\partial\theta_1}\mathcal{L} = \theta_1 + x_1\lambda = 0$$

$$\theta_1 = -\lambda x_1 = -\lambda a$$

$$\frac{\partial}{\partial\theta_2}\mathcal{L} = \theta_2 + x_2\lambda$$

$$\theta_2 = -\lambda x_2 = -\lambda e$$

- new Lagrangian

$$\mathcal{L} = \frac{1}{2}((-\lambda a)^2 + (-\lambda e)^2) + \lambda((-\lambda a)a + (-\lambda e)e) + \lambda$$

$$= \frac{1}{2}(\lambda^2 a^2 + \lambda^2 e^2) + -\lambda^2 a^2 - \lambda^2 e^2 + \lambda$$

$$\frac{\partial}{\partial\lambda}\mathcal{L} = a^2\lambda + e^2\lambda - 2a^2\lambda - 2e^2\lambda + 1$$

$$= -a^2\lambda - e^2\lambda + 1 = 0$$

$$1 = a^2\lambda + e^2\lambda \qquad \lambda = \frac{1}{(a^2 + e^2)}$$
$$= \lambda(a^2 + c^2)$$

- then

$$\theta_1 = -\lambda a = -\frac{a}{a^2+e^2}$$

$$\theta_2 = -\lambda e = -\frac{e}{a^2+e^2}$$

$$\boxed{\theta^* = \begin{bmatrix} -\frac{a}{a^2+e^2} \\ -\frac{e}{a^2+e^2} \end{bmatrix}}$$

b) $x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $c=1$

$y_1 = 1$ $y_2 = -1$

$$\mathcal{L}(\theta, \lambda_1, \lambda_2) = f(\theta) - \lambda(g_1(\theta) - c) - \lambda(g_2(\theta) - c)$$

$$= \frac{1}{2}\|\theta\|^2 - \lambda_1(y_1\theta^T x_1 - c) - \lambda_2(y_2\theta^T x_2 - c)$$

$$= \frac{1}{2}(\theta_1^2+\theta_2^2) - \lambda_1((\theta_1 x_{11} + \theta_2 x_{12})y_1 - 1)$$
$$- \lambda_2((\theta_1 x_{21} + \theta_2 x_{22})y_2 - 1)$$

$$= \frac{1}{2}(\theta_1^2+\theta_2^2) - \lambda_1(\theta_1 + \theta_2 - 1)$$
$$- \lambda_2(-\theta_1 - 1)$$

$$= \frac{1}{2}(\theta_1^2+\theta_2^2) - \lambda_1\theta_1 - \lambda_1\theta_2 + \lambda_1 + \lambda_2\theta_1 + \lambda_2$$

$=$

$$\frac{\partial}{\partial\theta_1}\mathcal{L} = \theta_1 - \lambda_1 + \lambda_2 = 0 \qquad \theta_1 = \lambda_1 - \lambda_2$$

$$\frac{\partial}{\partial\theta_2}\mathcal{L} = \theta_2 - \lambda_1 = 0 \qquad \theta_2 = \lambda_1$$

$$\mathcal{L}(\theta,\lambda_1,\lambda_2)=\frac{1}{2}\left((\lambda_1-\lambda_2)^2+\lambda_1^2\right)-\lambda_1^2+\lambda_1\lambda_2$$
$$-\lambda_1^2+\lambda_1-\lambda_2^2+\lambda_1\lambda_2+\lambda_2$$
$$=\frac{1}{2}\left(\lambda_1^2-2\lambda_1\lambda_2+\lambda_2^2+\lambda_1^2\right)-2\lambda_1^2+\lambda_1-\lambda_2^2+\lambda_2+2\lambda_1\lambda_2$$
$$=\lambda_1^2-\lambda_1\lambda_2+\frac{\lambda_2^2}{2}-2\lambda_1^2+\lambda_1-\lambda_2^2+\lambda_2+2\lambda_1\lambda_2$$
$$=-\lambda_1^2-\frac{1}{2}\lambda_2^2+\lambda_1\lambda_2+\lambda_1+\lambda_2$$

$$\frac{\partial}{\partial\lambda_1}\mathcal{L}=-2\lambda_1+\lambda_2+1=0 \quad ①$$
$$\frac{\partial}{\partial\lambda_2}\mathcal{L}=-\lambda_2+\lambda_1+1=0 \quad ②$$

$$\lambda_1-\lambda_2+1=0$$
$$2\lambda_1-2\lambda_2+2=0$$
$$-\lambda_2+3=0 \qquad \lambda_2=3$$

$$-3+\lambda_1+1=0 \qquad \lambda_1=2$$

$$\theta_1=2-3=-1, \quad \theta_2=2$$

$$\gamma=\frac{1}{\|\theta\|}=\frac{1}{\sqrt{\theta_1^2+\theta_2^2}}=\frac{1}{\sqrt{1+4}}=\frac{1}{\sqrt{5}}$$

$$\boxed{\theta^*=\begin{bmatrix}-1\\2\end{bmatrix}, \quad \gamma=\frac{1}{\sqrt{5}}}$$

c) $\frac{1}{2}\|\theta\|^2$ $\quad y_n(\theta^T x_n + b) \geq 1$ $\quad x_1 = \begin{bmatrix}1\\1\end{bmatrix}, y_1 = 1$

$\uparrow f(\theta)$ $\quad \uparrow g(\theta)$ $\quad x_2 = \begin{bmatrix}1\\0\end{bmatrix}, y_2 = -1$

$$\mathcal{L}(\theta, b, \lambda) = f(\theta) - \lambda_1 (g_1(\theta) - c) - \lambda_2 (g_2(\theta) - c)$$
$$= \frac{1}{2}\|\theta\|^2 - \lambda_1 (y_n(\theta^T x_n + b) - c)$$
$$- \lambda_2 (y_n(\theta^T x_n + b) - c)$$
$$= \frac{1}{2}(\theta_1^2 + \theta_2^2) - \lambda_1 (y_n(\theta_1 x_{11} + \theta_2 x_{12} + b) - c)$$
$$- \lambda_2 (y_n(\theta_1 x_{21} + \theta_2 x_{22} + b) - c)$$
$$= \frac{1}{2}(\theta_1^2 + \theta_2^2) - \lambda_1 (\theta_1 + \theta_2 + b - 1)$$
$$- \lambda_2 (-\theta_1 - b - 1)$$

$$\frac{\partial}{\partial \theta_1}\mathcal{L} = \theta_1 - \lambda_1 + \lambda_2 = 0, \quad \theta_1 = \lambda_1 - \lambda_2$$

$$\frac{\partial}{\partial \theta_2}\mathcal{L} = \theta_2 - \lambda_1 = 0, \quad \theta_2 = \lambda_1$$

$$\frac{\partial}{\partial b}\mathcal{L} = -\lambda_1 + \lambda_2 = 0, \quad \lambda_1 = \lambda_2$$

$$\mathcal{L}(b, \lambda) = \frac{1}{2}((\lambda_1 - \lambda_2)^2 + (\lambda_1)^2)$$
$$- \lambda_1 (\lambda_1 - \lambda_2 + \lambda_1 + b - 1)$$
$$- \lambda_2 (-\lambda_1 + \lambda_2 - b - 1)$$
$$= \frac{1}{2}(2\lambda_1^2 - 2\lambda_1\lambda_2 + \lambda_2^2) - 2\lambda_1^2 + \lambda_1\lambda_2 - \lambda_1 b + \lambda_1$$
$$+ \lambda_1\lambda_2 - \lambda_2^2 + b\lambda_2 + \lambda_2$$
$$= \lambda_1^2 - \cancel{\lambda_1\lambda_2} + \frac{1}{2}\lambda_2^2 - 2\lambda_1^2 + \lambda_1\lambda_2 - \lambda_1 b + \lambda_1$$
$$+ \cancel{\lambda_1\lambda_2} - \lambda_2^2 + b\lambda_2 + \lambda_2$$
$$= -\lambda_1^2 - \frac{1}{2}\lambda_2^2 + \lambda_1\lambda_2 + b\lambda_2 - b\lambda_1 + \lambda_1 + \lambda_2$$

$\lambda_1 = \lambda_2$

$$\mathcal{L}(\lambda) = \cancel{-\lambda^2} - \frac{1}{2}\lambda^2 \cancel{+\lambda^2} + \cancel{b\lambda} \cancel{- b\lambda} + 2\lambda$$
$$= -\frac{1}{2}\lambda^2 + 2\lambda$$

$$\frac{\partial}{\partial \lambda}\mathcal{L} = -\lambda + 2 = 0, \quad \lambda = 2$$

$\lambda_1 = 2, \quad \lambda_2 = 2$

- substitute back

$$\theta_1 = \lambda_1 - \lambda_2 = 0$$
$$\theta_2 = \lambda_1 = 2$$

$$y_n(\theta^T x_n + b) \geq 1$$

① $y_1(\theta_1 x_{11} + \theta_2 x_{22} + b) \geq 1$

$1(0 \cdot 1 + 2 \cdot 1 + b) \geq 1$

$2 + b \geq 1$

$b \geq -1$

② $y_2(\theta_1 x_{21} + \theta_2 x_{22} + b) \geq 1$

$-1(0 \cdot 1 + 2 \cdot 0 + b) \geq 1$

$-b \geq 1$

$b \leq -1$

$b \geq -1$ and $b \leq -1 \Rightarrow b = -1$

$$\gamma = \frac{1}{\|\theta\|} = \frac{1}{\sqrt{\theta_1^2 + \theta_2^2}} = \frac{1}{\sqrt{4}} = \frac{1}{2}$$

$$\boxed{\theta^* = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \quad \gamma = \frac{1}{2}}$$

The margin γ with offset is larger than without offset.

3.1
a)
done

b)
done

c)
done

3.2
a)
done

b)
It makes sense to maintain class proportions across folds during cross validation in order to have the training set match the original dataset as closely as possible. If we did not maintain class proportions, it’s possible that the training set may mostly only contain data from one class. As a result, the model will not be very accurate.

c) done

d)

C	**Accuracy**	**F1-Score**	**AUROC**	**Precision**	**Sensitivity**	**Specificity**
10^{-3}	0.7089	0.8297	0.5000	0.7089	1.000	0.000
10^{-2}	0.7107	0.8306	0.5031	0.7102	1.000	0.0063
10^{-1}	0.8060	0.8755	0.7188	0.8357	0.9294	0.5081
1	0.8146	0.8749	0.7531	0.8562	0.9017	0.6045
10	0.8182	0.8766	0.7592	0.8595	0.9017	0.6167
10^{2}	0.8182	0.8766	0.7592	0.8595	0.9017	0.6167
best C	10	10	10	10	10^{-3}	10

For accuracy, F1-score, AUROC, precision, and specificity, the cross validation performance has a direct relationship with the value of C. As the value of C increases, the performance score also increases. For sensitivity, the cross validation performance has an inverse relationship with the value of C. As the value of C increases, the performance score decreases.

3.3

a)

Gamma is a hyperparameter that is only present in the SVM's rbf kernel. Gamma represents how much influence a particular data point reaches. If the gamma value is too high, data points closer to the decision boundary carry more influence. This means that our variance is low and the model will be more likely to overfit. If the gamma value is too low, data points closer to the decision boundary will carry similar influences as those farther away from the decision boundary. This means that our variance is high and the model will be more likely to underfit.

b)

We first use the same ranges for C and gamma as the linear kernel. This gives us a baseline of what the best C and gamma values are. We find that a good value of C is 100, and a good value of gamma is 0.01. We then adjust the ranges of C and gamma with the good values that we just found as a centroid. We repeat this process until we find the optimal values for C and gamma. We find that the optimal value of C is 80, and the optimal value of gamma is 0.005.

c)

Metric	**Score**	**C**	**Gamma**
Accuracy	0.8218	80	0.005
F1 Score	0.8800	80	0.005
AUROC	0.7618	80	0.005
Precision	0.8621	80	0.005
Sensitivity	0.9067	70	0.004
Specificity	0.6169	70	0.004

It seems that performance is directly proportional to C. As C increases, so does performance. However, performance is quadratically proportional to gamma. As gamma increases, performance increases up to a certain point and then decreases.

3.4

a)

For the SVM with linear kernel, we'll pick C to be 100. Performances with C values around this value are the highest. For the SVM with rbf kernel, we'll pick C to be 80 and gamma to be 0.005. As before, performances with the above hyperparameters perform the best.

b)
done

c)

Metric	**Linear Kernel**	**RBF Kernel**
Accuracy	0.7429	0.7429
F1 Score	0.4375	0.4375
AUROC	0.6259	0.6259
Precision	0.6364	0.6364
Sensitivity	0.3333	0.3333
Specificity	0.9184	0.9184

It appears that both types of kernels performed exactly the same on all of the metrics. This is a surprising result, because we would typically expect a particular type of kernel to perform better on a particular dataset. In this case, perhaps the data is separable such that both linear and rbf SVM models classify them in the exact same way, thus yielding the same scores.